(U)

ANALYSIS OF DNA SEQUENCES BY AN OPTICAL TIME-INTEGRATING CORRELATOR: PROPOSAL (U)

by

N. Brousseau and R. Brousseau

DEFENCE RESEARCH ESTABLISHMENT OTTAWA
TECHNICAL NOTE 91-33

Canada

November 1991
Ottawa

National Defence / Défense nationale

ANALYSIS OF DNA SEQUENCES BY AN OPTICAL TIME-INTEGRATING CORRELATOR: PROPOSAL (U)

by

N. Brousseau
Communications Electronic Warfare Section
Electronic Warfare Division

and

R. Brousseau
Institut de Recherche Biotechnologie
6100 au Royalmount
Montréal, Québec
H4P 2R2

DEFENCE RESEARCH ESTABLISHMENT OTTAWA
TECHNICAL NOTE 91-33

PCN
041LK11

November 1991
Ottawa

ABSTRACT

This technical note presents a proposal to perform the analysis of DNA sequences with an analogue optical computer. The DNA analysis involves the computation of a massive amount of correlations. A time-integrating correlator is an ideal tool to perform that processing at a very fast speed. A design based on commercially available equipment is presented together with a comparison of the processing time of the system with conventional computer technology. The speed of this design is orders of magnitude greater than existing techniques. An overview of the technology already available for such a project is presented together with an outline of the areas that need more development.

RESUMÉ

Cette note technique propose un système d'analyse des séquences d'ADN par un ordinateur analogique optique. L'analyse de séquences d'ADN implique le calcul d'une énorme quantité de corrélations. Un corrélateur à intégration temporelle est un outil idéal pour effectuer rapidement ce type d'opération. On présente un système conçu à partir d'équipement commercialement disponible ainsi qu'une comparaison entre ses temps de traitement et les temps de traitement d'ordinateurs conventionnels. L'amélioration de la vitesse de traitement est de plusieurs ordres de grandeur et est suffisante pour permettre d'envisager le traitement de tout le génome humain. On présente finalement une revue de la technologie déjà disponible pour la construction du système ainsi qu'un aperçu des domaines necessitant encore du développement.

EXECUTIVE SUMMARY

Molecular biologists tell us that each cell in our body carries all the information necessary to reconstruct the entire organism. This information is stored in a molecular structure called DNA and the analysis of DNA sequences is of particular interest for the understanding of the basic processes governing life. In that context, the mission of the Human Genome Project is to map the entire mosaic of the human DNA. In an effort to reach that objective, biochemists try to match a particular segment of DNA to existing data banks, with the possibility that the match will not be perfect. Correlation techniques implemented on digital computers are used to perform the analysis on the limited amount of data available today and the process is tedious. Considering that only a small fraction of the $3x10^9$ human genome nucleotides is now available in the data banks, a mapping of the entire human genome requires a computational breakthrough.

This technical note proposes a new method to perform the analysis of human or animal DNA sequences with an analogue optical computer. The new method is characterized by short processing times that make the analysis of the entire human genome a tractable enterprise. The proposal is based on the utilization of a Time-Integrating Correlator (TIC). This type of optical correlator is particularly well suited to the very fast correlation of long data streams such as the data involved in the analysis of DNA. A design based on commercially available equipment is presented together with a comparison of the processing time of the system with conventional computer technology. Comparison of the expected processing times of a TIC, for a particular case, leads to the conclusion that the TIC could be 10 times faster than a 80 Mega Instructions Per Second (MIPS) computer and over 375 times faster than a personal computer. An overview of the technology already available for such a project and an outline of the areas that need further development is also included.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (cont.)

LIST OF TABLES

LIST OF ABBREVIATIONS

DNA:	deoxyribonucleic acid
TIC:	time-integrating correlator

LIST OF VARIABLES

t:	time
A(t):	signal applied to the Bragg cell A
B(t):	signal applied to the Bragg cell B
T:	integration time
v:	propagation velocity of the acoustic waves in the Bragg cells
S(T,z):	signal produced by the detector array
z:	distance along the Bragg cells or their images
D:	bit rate
τ:	transit time of the signal in the Bragg cell
2τ:	time-delay window of the correlator
n:	number of bases in the query sequence
r:	repetition of the bits
R:	length of the representation
d_s :	duration of the query sequence

1.0 INTRODUCTION

Optical information processing has been developing since the early 1960's, first slowly, then at an accelerated pace. It is now a field of intense activity. Although the situation is likely to change quickly from now on, analogue optical computing has been the area of optical processing that has had the most success in terms of the development of practical systems. Acousto-optic spectrum analyzers for the processing of wideband military radar signals and synthetic aperture radar correlators are probably among the best examples of analogue optical processors dedicated to specific applications. However, the achievement of a truly significant breakthrough in optical computing has been elusive [1].

In this paper, we are proposing the application of a late 1970's concept to the solution of a 1990's problem. The problem considered here is the analysis of human or animal DNA sequences where biochemists attempt to match a query sequence of DNA to an identical or a similar segment that may be present in the existing computer databases. The genome of a particular living organism is all its genetic information that is encoded in DNA sequences. In this context, the mission of the Human Genome Project [2-4] is to map and sequence the entire mosaic of the human DNA. Correlation techniques implemented on digital computers are used to do the sequence matching on the limited amount of data available today and the process is tedious. Considering that only a small fraction of the 3×10^9 human genome nucleotides is now available and stored in the data banks, a computational breakthrough is required to allow the processing of the entire human genome.

The solution that we are proposing for the analysis of DNA sequences is to use a Time-Integrating Correlator (TIC) whose theory and architecture were studied in the late 1970's and the 1980's. This type of optical correlator is particularly well suited to the very fast correlation of long data streams such as the data involved in the analysis of DNA. The limitations on dynamic range that are a problem in the application of analogue computing to noisy radar or communication signal processing, are not a problem here. The data to be correlated comes from a computer database and is noiseless.

DNA sequences are built from four bases represented by the letters A, C, G and T. A fifth letter ,N, is used to represent unknown elements at particular locations in a sequence. Optical approaches have already been considered for the analysis of DNA sequences. Vander Lugt space integrating correlators have been proposed [5,6]. In this approach, the DNA bases and the sequence they form are represented by two dimensional arrays on which pattern recognition is performed with a Vander Lugt space integrating correlator. The advantages of using the TIC approach are that there is no need to transform the one-dimensional DNA sequences into two-dimensional patterns and the complex process of generating and changing matched filters is eliminated.

In an operational system, it is proposed that the TIC would be an optical black box performing rapid correlations under the control of a computer. The TIC would be a high performance correlation module integrated to a software environment already familiar to the users. The crucial difference would be the increased speed of operation of the system. In this respect, the proposed system meets the concerns for gradual insertion of optical technology[7] into information processing systems. This is viewed as necessary for the progress of optical computing in the 1990's and beyond.

2.0 SCOPE OF THE PROBLEM AND CURRENT TECHNOLOGY

All living organisms encode their genetic information in the same way, by using linear polymers of phosphoric acid and sugar (deoxyribose) upon which are attached four different bases, adenine (A), cytosine (C), guanine (G) and thymine (T). These linear polymers of very long extent (one chromosome of 4×10^6 units for a typical bacterium, twenty-three chromosomes of up to 200×10^6 units for human beings) contain regions called genes, which are translated into proteins, as well as regulatory regions and regions of as yet unknown function. These linear polymers can be read sequentially by chemical and enzymatic techniques and the resulting linear information interpreted as to their function. Of particular interest in the human genome are regions responsible for genetics defects; once these are located and identified, then early treatment may become possible.

Over the past ten year, DNA sequencing techniques have advanced sufficiently for a modest start to be made on harvesting and analyzing the formidable array of genetic diversity in life forms. Most of the DNA sequence information available today is tabulated in the GenBank* database. Release 65 (September 1990) of this database contains 49×10^6 nucleotides from all organisms, divided into thirteen divisions. The Primate division, where human sequence data is located, accounts for 8×10^6 nucleotides. While the amount of information harvested so far is very small compared even to a single human genome, it is already apparent that present day computer technology will be unable to deal with future developments. A complete search of GenBank Release 65 with a query sequence of 3000 bases currently takes about 8 minutes of CPU time on a 80 MIPS mainframe computer or over five hours on a personal computer operating at 25 MHz. As the database grows towards its projected size of 3×10^9 for the human genome alone (discounting inevitable overlaps, repetitions and person-to-person variations), it can be foreseen that current equipment will quickly become utterly impractical to use.

* Produced by GenBank c/o IntelliGenetics Inc. 700 East El Camino Real, Mountain View CA 94040.

3.0 TIME-INTEGRATING CORRELATORS

Time-Integrating Correlators are analogue optical computers designed to perform the correlation of two signals. The many possible ways to build TICs are well documented in the literature [8-27] and good review papers are available [15,21,27]. The various factors affecting the performance of these systems have also been extensively studied [28-33]. We present in this section a brief review of the principle of operation of TICs with emphasis on the characteristics and parameters that have an impact on the design and operation of a TIC applied towards the analysis of DNA sequences.

We have chosen to illustrate the concept of a TIC using the Mach-Zehnder architecture (see Figure 1). Other architectures may have distinct implementation advantages, such as compactness [17-19], but it is easier to explain the principle of operation of the TICs by using the Mach-Zehnder configuration as a model.

The first operation performed by a TIC is the transformation of the electrical signals A and B to be correlated into modulated light beams. The two input signals are applied to piezoelectric transducers attached to the Bragg cell crystals and acoustic waves are generated. They propagate in the Bragg cell crystals (see Figure 2) thus forming a moving grating of changing indices of refraction. The two Bragg cells are then illuminated by expanded laser beams and the laser light interacts with the acoustic waves. Some of the incident light is diffracted through the acousto-optic interaction. The information contained in the electrical signals applied to the Bragg cells is thus transferred to the diffracted laser beams. The relative positions of the Bragg cells and of the illuminating beams are arranged to produce two diffracted beams that are mixed together with a half-silvered mirror. The signals propagating in the Bragg cells are imaged, with a lens, onto a linear array of light detectors in such a way as to be counterpropagating. The detector array performs a coherent addition and a time-integration of the two images. Detailed analysis of the signal produced by the detector array are available in the literature [8,18,21,27,30-33].

We will use here simplified expressions that emphasize the ability of the system to produce the correlation of the two input signals A(t) and B(t). The distance along the Bragg cells and their images is described by the variable z and the origin, $z=0$, which is defined to be at the centre of the Bragg cells and correspondingly, at the centre of their images on the detector array (see Figure 1). If A(t) and B(t) are the signals applied to the Bragg cells, T is the integration time of the detector array and v is the velocity of propagation of the acoustic signal in the Bragg cells, the signal S(T,z) produced by the detector array can be described by:

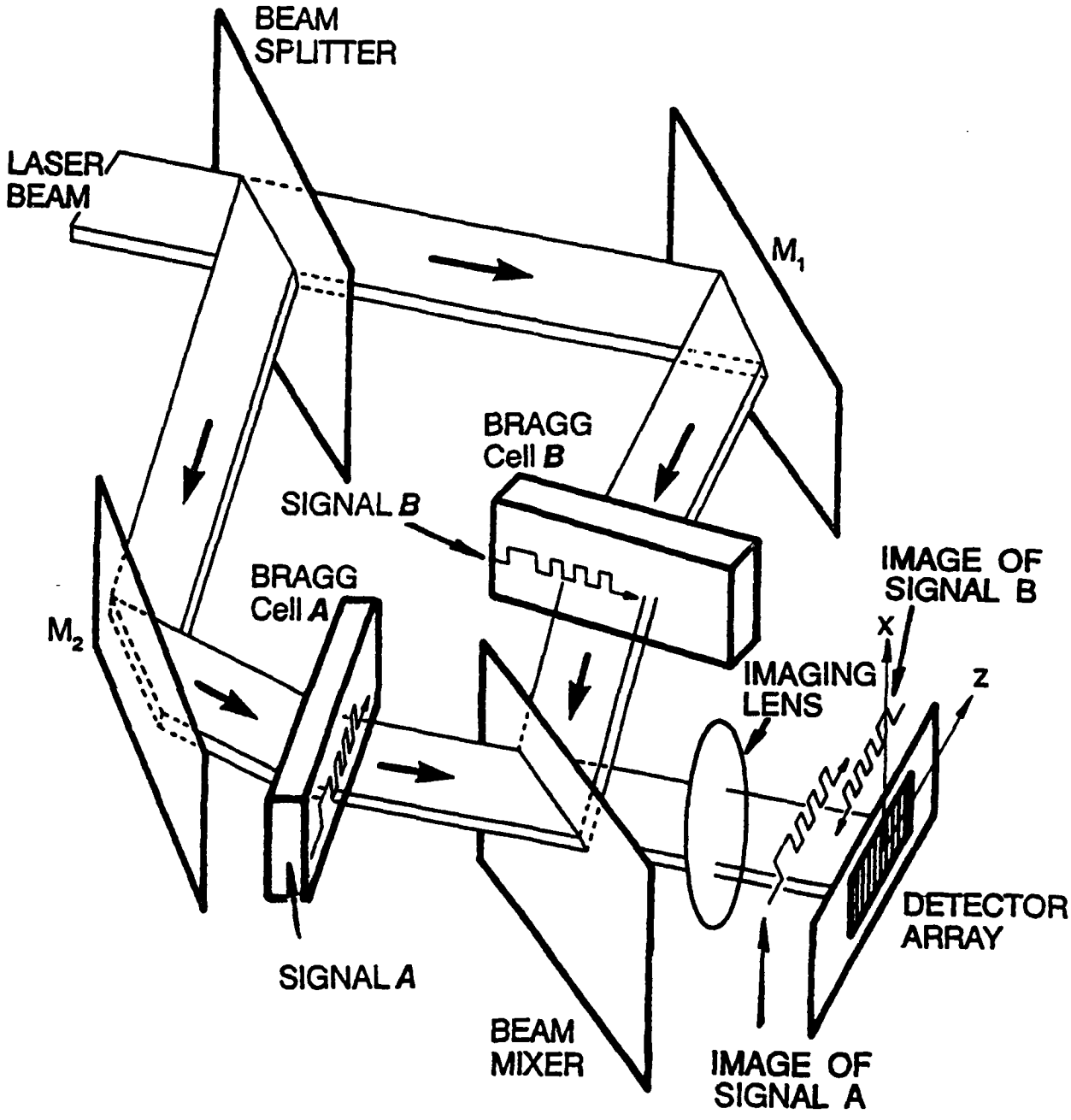

Figure 1: Time-integrating correlator: Mach-Zehnder architecture. The beam splitter separates the incident laser beam into two paths. M1 and M2 are folding mirrors. The two beams diffracted by the Bragg cells are mixed together by a beam mixer. The two diffracted light distributions are coaxial and imaged in such a way as to be counterpropagating on the detector array that performs a time-integration.

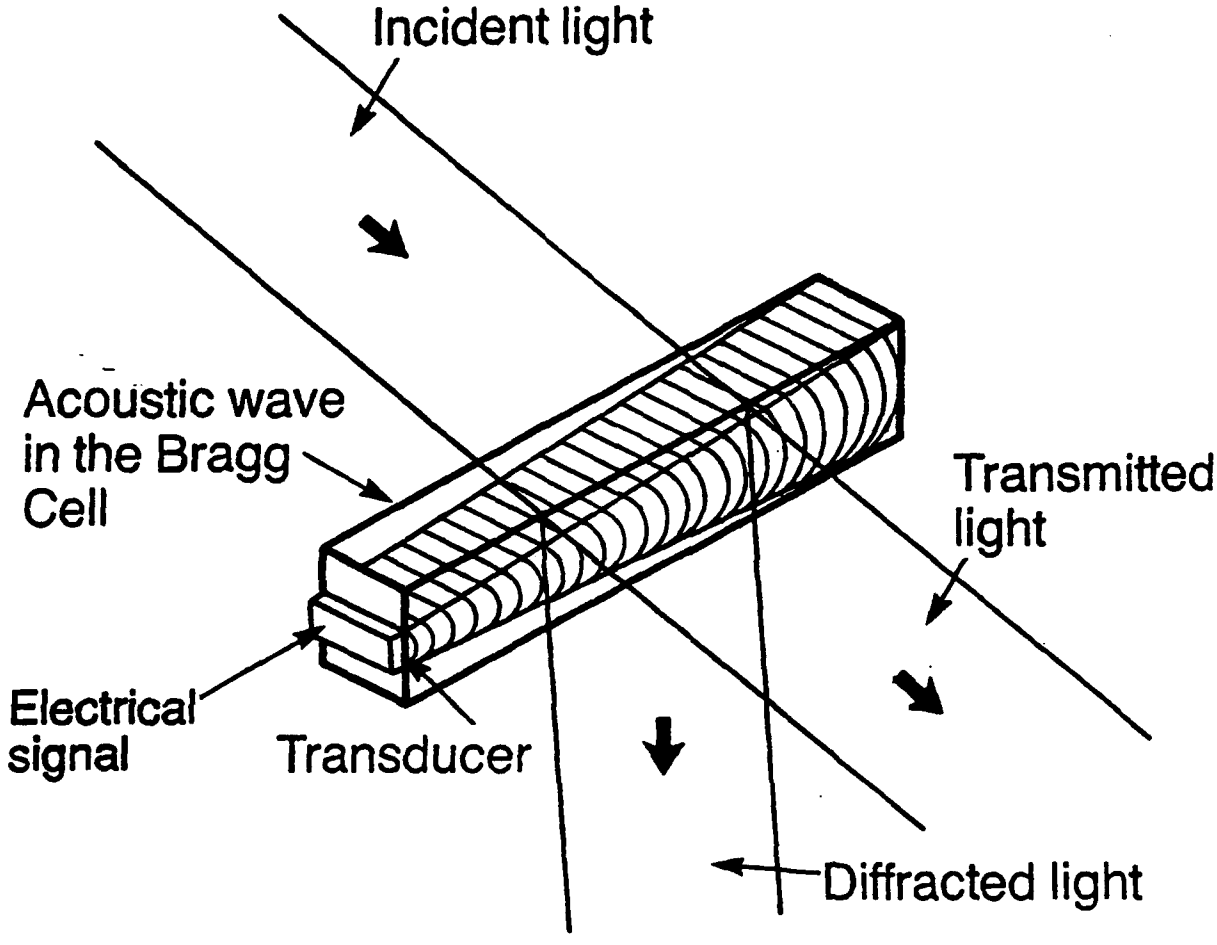

Figure 2: Bragg cell operation. The electrical input is applied to the piezoelectric transducer that generates a moving grating of changing indices of refraction. That moving grating diffracts some of the light illuminating the Bragg cells and the information contained in the electrical input is transferred to the diffracted laser beam.

$$S(T,z) = \int_T |A(t+z/v) + B(t-z/v)|^2 \, dt \quad (1)$$

$$S(T,z) = \int_T A^2(t+z/v)dt + \int_T B^2(t-z/v)\, dt + 2\int_T A(t+z/v)B(t-z/v)\, dt \quad (2)$$

The first two terms of equation 2 correspond to a pedestal on which rides the correlation peak formed by the AB term (see Figure 3). The presence of a peak indicates that the two input signals A and B are identical and conversely the absence of the correlation peak indicates that the two inputs are different. The presence of a correlation peak with a reduced height indicates that A and B have similarities but are not identical.

One of the most interesting characteristics of the TIC is that the correlation peak is produced at the meeting point of the images of the two counterpropagating signals (see Figure 1) on the detector array. If τ is the transit time of the signals in the Bragg cells, the width of the time-delay window over which it is possible to observe the peak is 2τ because the signals are counterpropagating on the detector array. If the signal duration is longer than the time-delay window, and if the difference of time of arrival of the signals is such that the meeting point is outside the time-delay window, then no correlation peak will be observed. It will then be necessary to try different time-shifts of one of the signals to move the correlation peak into the time-delay window of the TIC. The time-shifts should be designed to produce contiguous or slightly overlapping time-windows. If bits are defined as the 0 and 1's of the signals applied to the Bragg cells, the number of time-delay steps (measured in number of bits) that can be processed within one time-delay window of the TIC is given by the duration of the time-delay window 2τ multiplied by the bit rate of the input signals, D,

$$2\tau \times D. \quad (3)$$

It is also desirable to emphasize the correlation peak by removing the pedestal shown in Figure 3. The technique used here involves subtraction of two successive frames collected by the detector with a 180° phase shift on one of the signals applied to the Bragg cells for the collection of the second frame[8,21]. The effective integration time generated by the phase-shift method is twice the integration time of the detector array. Another technique uses a reference frame that has a pedestal but does not contain a peak. It is regularly updated and subtracted from every new frame that is collected. This pedestal removal technique produces effective integration times that are equal to the integration time of the detector array but it has to be used in conjunction with a frame that always contains a positive peak.

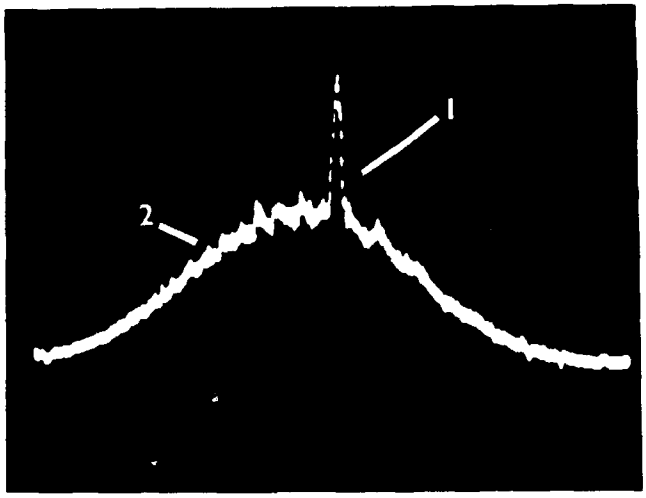

Figure 3: Typical output from a TIC: (1)-correlation peak formed by the AxB term and (2)-pedestal formed by the A^2 + B^2 terms.

Most aspects of the operation of the TIC are controlled by a computer that we will call the Controller. The Controller determines the input of the data to the TIC and the collection of data from the detector array, including pedestal removal and peak detection.

4.0 REPRESENTATIONS OF THE DNA BASES

DNA sequences are built from four bases represented by the letters A, C, G and T. A fifth letter, N, is used to represent unknown elements at particular locations in a sequence. The sequences representing segments of the human genome have to be transformed into electrical signals suitable as inputs to the Bragg cells (see Figure 2). One way to accomplish this is to represent each base by a binary pseudorandom sequence of the type used in spread spectrum code division multiple access communications [34, chap.3]. The bits (0 and 1's) specified by the representations of the bases can be implemented using binary phase-shift-keyed modulation [34, p.16-18]. The short representations listed in Table 1 and Figure 4 which have been selected for the low value of their cross-correlation could be used.

Table 1: Short representations of the dna bases where each base is represented by 7-bits long pseudorandom sequences

Adenine (A)	0 0 0 0 0 0 1
Cytosine (C)	0 1 0 0 1 1 0
Guanine (G)	1 0 1 0 0 1 0
Thymine (T)	1 1 0 1 0 0 0
Unknown (N)	1 1 1 0 1 0 1

Figure 5 represents the flow of data in a DNA analysis system based on an optical TIC. On the left side, the human genome data base has a potential of 3 billion bases. Currently there are approximately 50 million bases of sequence available from all living organisms. The 50 million bases that are known are stored in a digital database where they are designated by letters. These letters are then represented by pseudorandom binary sequences and transformed into analogue signals which are suitable to operate a Bragg cell. The right side represents the new query sequence acquired by a scientist. It undergoes the same transformation and is correlated with the database from the left side by the TIC. The results are displayed and if the query sequence was not already included in the known DNA database, it is incorporated into the database.

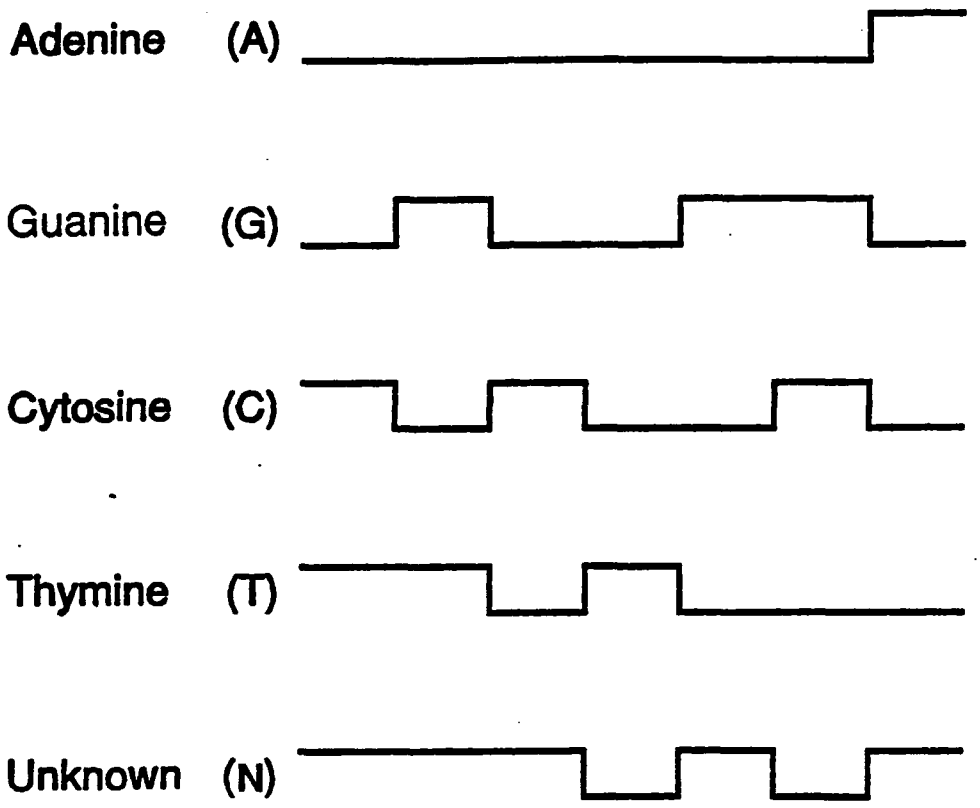

Figure 4: Short representations of the DNA bases where each base is represented by a 7-bits long pseudorandom sequence.

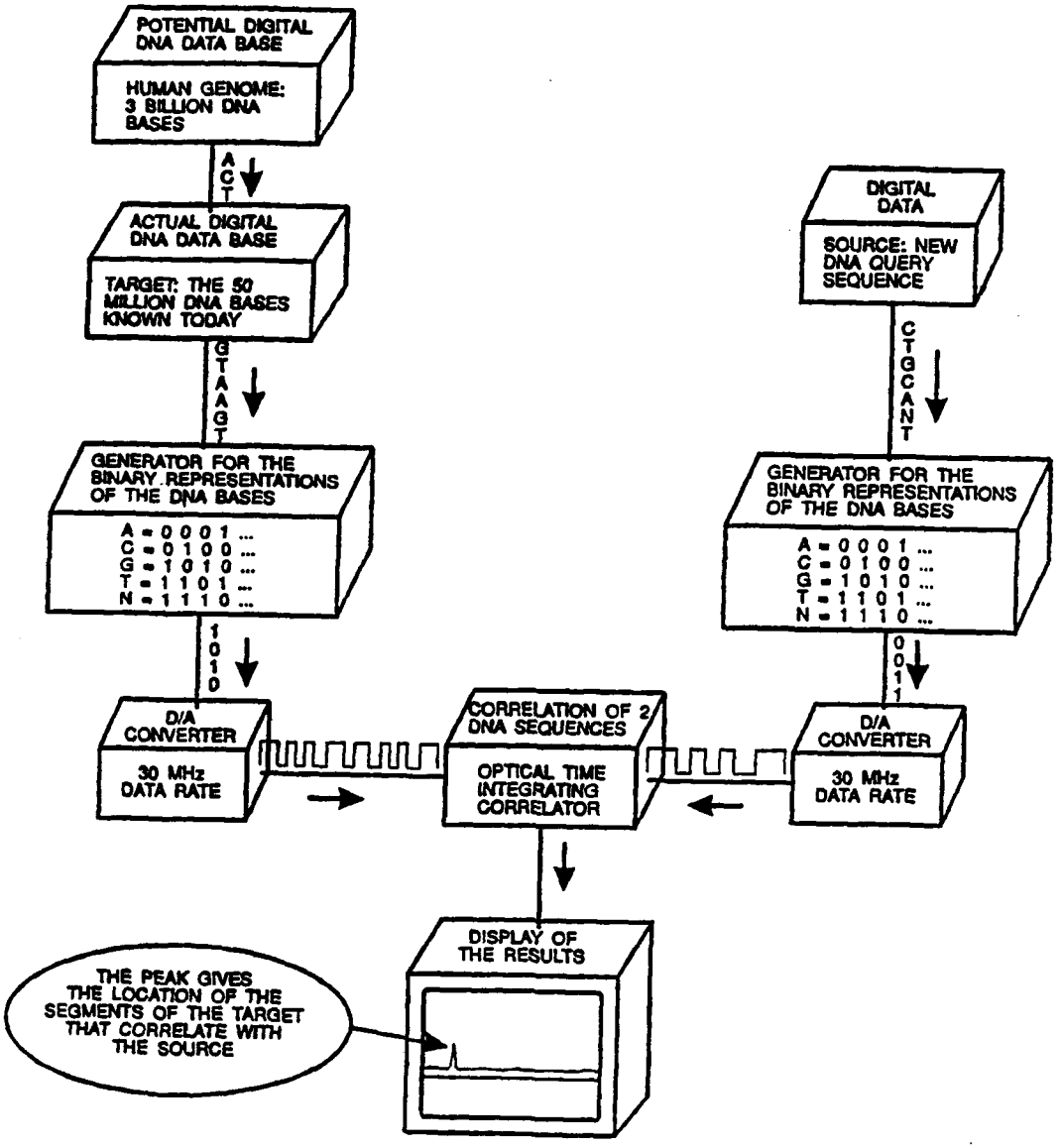

Figure 5: The flow of data in a DNA analysis system based on an optical TIC. On the left side the human genome has a potential of 3 billion bases. The 50 million bases that are known are stored in a digital database where they are designated by letters. These letters are then represented by pseudorandom binary sequences and transformed into analogue signals which are suitable to operate a Bragg cell. The right side represent the new data (query sequence) acquired by a scientist. It undergoes the same transformation and is correlated by the TIC with the data from the database on the left side. The results are displayed and if the query sequence was not already included in the known DNA data base, it is incorporated.

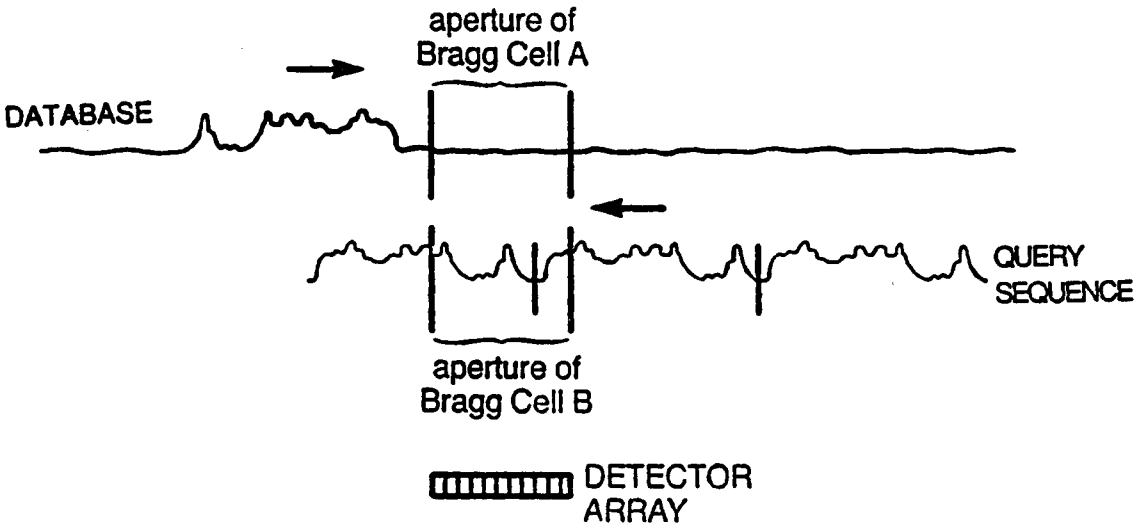

Figure 6: Coarse analysis of a DNA sequence. A database is illustrated as it propagates through Bragg cell A just before the passage of the segment that is identical to the query sequence. The signal formed by the repetitions of the query sequence is illustrated at the same moment in Bragg cell B. The correlation peak will start formation a few moments later, in about the transit time in the Bragg cell divided by two.

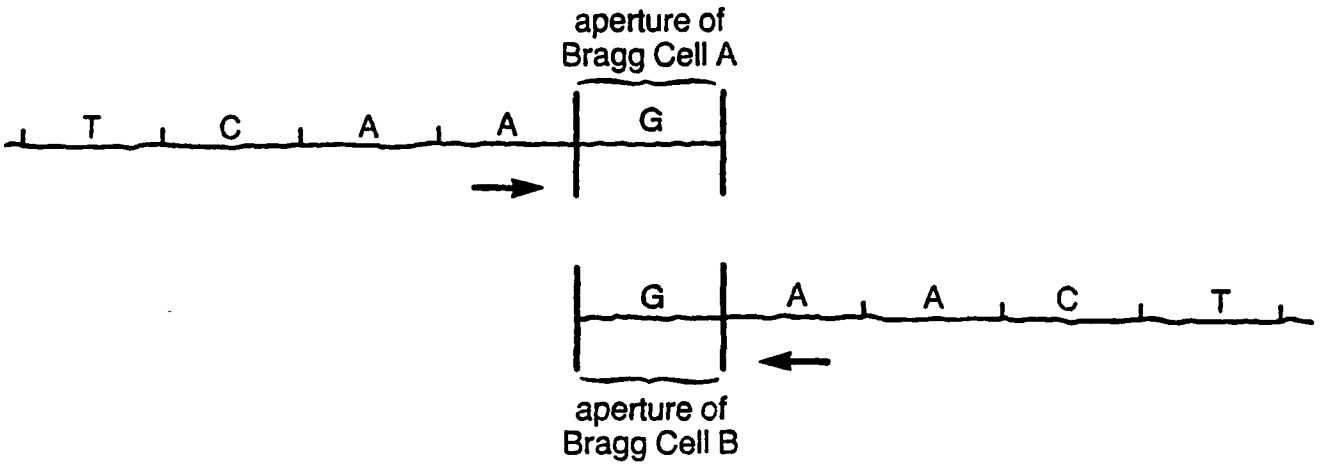

Figure 7: Fine, base-by-base analysis of a DNA sequence. The database and the query sequence are represented by long pseudorandom sequences that almost fill the Bragg cells' apertures. The system is illustrated at the moment when the base G is correlating.

5.0 DNA ANALYSIS STRATEGY

The purpose of this section is to present a strategy to implement the analysis of a DNA sequence with a TIC. We wish to find segments of the database that are identical or similar to the query sequence and their location within the database. We also want to produce a base-by-base comparison of the query sequence using the segments of the database that are identified as correlating with the query sequence. The analysis is made using a two-level procedure. A coarse analysis is first used to locate the areas of the database that are similar or identical to the query sequence (see Figure 6). Then, a fine analysis (see Figure 7), is performed on the database segments identified by the coarse analysis to establish the map of conformity.

6.0 STRATEGY FOR COARSE ANALYSIS

6.1 Introduction

The purpose of the coarse analysis is to find the areas of the database that are similar to the query sequence. The process involved in the production of the correlation peaks for the coarse analysis consists of sending the database sequence without interruption through Bragg cell A (see Figure 6). Simultaneously, the query sequence is passed through Bragg cell B continuously. The output of the detector array is examined at regular intervals T. The pedestal is removed and the presence of a peak is verified by comparison with a preset threshold level for each collected frame. The setting of the threshold level determines the degree of similarity that is required to declare that a certain segment of the database correlates with the query sequence. The higher the peak, the better the correlation between the query sequence and the database. These operations can be performed in real time with a proper hardware implementation. When a segment of the database in Bragg cell A is identical or sufficiently similar to the query sequence in Bragg cell B, correlation peaks will be produced and detected. The time of occurrence of such events is associated with the position of the query sequence in the database and can be determined by knowing which frame contains the correlation. All the occurrences of a correlation peak will be noted and the fine analysis will follow to obtain a base-by-base comparison of the query sequence with the database.

The length of the query sequences used can be very different. A system that would handle query sequences containing between 12 and 10^5 bases would be considered a valuable research tool by biochemists. Although only approximately 50×10^6 bases of DNA sequence are presently identified from all living species, the system should be able to handle the full human genome of 3×10^9 bases. Our design of DNA analysis with an optical TIC is based on these numbers.

6.2 Query Sequence Duration Issues

The appearance of the correlation peak is usually not synchronized with the beginning of the integration periods. For example, a peak could start formation halfway through an integration period. To ensure that at least one integration period produces a peak of maximum height, the duration of the query sequence d_s must be at least twice the detector integration time T. If we assume that the pedestal removal is done by subtracting a reference frame, we must then have

$$d_s > 2T \tag{4}$$

If the phase shift pedestal removal technique (see section III) was used, it would be $d_s>4T$ because the correlation function is obtained from the subtraction of two successive frames with an effective integration time of 2T.

When a query sequence duration is short, it has to be stretched to meet this criteria. One approach is to use longer representations of the DNA bases. Another approach is to repeat each bit enough times to extend sufficiently the query sequence duration. If representations with bit repetition are chosen for the bases of the query sequence, the same representations have naturally to be used for the bases of the database. Stretching the query sequence duration can also be achieved by reducing the bit rate. Whatever stretching method is used, longer analysis times will result.

It is desired that the query sequence duration d_s be greater than 2T, to ensure the formation of at least one maximum height peak. However, d_s should also be less than 2τ, the time-delay window of the TIC, to avoid having to bring the correlation peak within the time-delay window of the TIC using different time-shifts to explore all possible relative delays between the query sequence and the database.

6.3 Analysis Times

Table 2 lists the analysis times, as a function of the number of bases in the query sequence, associated to one of the many possible strategies that combine bit repetition with bit rate adjustments for the analysis of a $50x10^6$ bases database corresponding to the DNA information available today. Representations of length 7 bits have been used (see Table 1).

For the sake of discussion, an integration time T of 50 μs (corresponding to a 1000-element detector array with a 20 MHz read-out frequency) and a time-delay window, 2τ, of 200 μs (corresponding to a 100 μs aperture Bragg cell) were selected. These numbers are representative specifications of commercial equipment currently available. Table 2 was compiled by adjusting the bit repetition and the bit rate to obtain appropriate query

Table 2: Analysis time for a $50x10^6$ bases database as a function of the number of bases in the query sequence. Query sequence lengths between 12 and 100000 are illustrated. The analysis time for query sequences longer than 857 bases grows linearly as the length of the query sequence because, for a longer query sequence, more time-shifts are required to find the correlation peak for a longer query sequence.

Query Sequence Length (Number of Bases)	Bit Repetition	Bit Rate (MHz)	Duration of Query Sequence (μs)	Analysis Time for One Phase (seconds)	Number of Time Shifts	Total Analysis Time (seconds)
12 to 14	2	1	168 to 196	700	1	700
15 to 28	1	1	105 to 196	350	1	350
29 to 42	1	2	102 to 147	175	1	175
43 to 57	1	3	100 to 133	116	1	116
58 to 71	1	4	101 to 124	86	1	86
72 to 142	1	5	101 to 199	70	1	70
143 to 214	1	10	100 to 150	35	1	35
215 to 285	1	15	100 to 133	23	1	23
286 to 428	1	20	100 to 150	18	1	18
429 to 10^5	1	30	100 to 23333	12	1 to 117	12 to 1404

sequence durations. The query sequence lengths used ranged from 100 μs (twice the integration time T) to 200 μs (the time-delay window of the TIC) for the reasons given in Section 6.2. The transition from dual to single bit representation, and to higher bit rates is made for the shortest query sequence length possible to obtain minimum processing times. Only one TIC was assumed to be available to do the processing. The data in Table 2 is plotted in Figure 8 where the left side uses a log-log axis and covers query sequences of length 12 to 857. A semi-log axis is more convenient for the right side because the analysis time varies linearly with the length of the query sequence. The abscissa and the ordinate are drawn on a logarithmic and a linear scale, respectively.

6.4 Examples

Examples drawn from Table 2 are discussed in detail in the following paragraphs. Let us assume that we do not want to use a bit rate less than 1 MHz and that the database contains 50×10^6 bases. For an integration time T of 50 μs, the query sequence duration d_s is:

$$d_s = n R r / D, \qquad (5)$$

where n is the number of bases in the query sequence, R is the length of the representations, r is the repetition of the bits and D is the bit rate. To ensure at least one frame with a full height correlation peak the query sequence duration d_s should therefore be at least 100 μs.

1) Let us consider a query sequence that contains 12 bases (see Table 2). If a seven bit long representation is used with a repetition of two and a bit rate of 1 MHz, the query sequence has a duration of 168 μs. The database has to be generated with the same parameters. A database made of 50×10^6 bases will have a duration d_t of 700 seconds. The 200 μs time-delay window of the TIC is sufficient to see the whole query sequence, so only one time-shift has to be tried and one 700 second run is required to analyse the data.

2) It is possible to analyse query sequences that are between 12 and 14-bases long with the same parameters because a 14-bases long query sequence has a duration of 196 μs. However, it is advantageous to reduce the analysis time by introducing a lower number of repetitions and by increasing the bit rate as the number of bases in the query sequence increases.

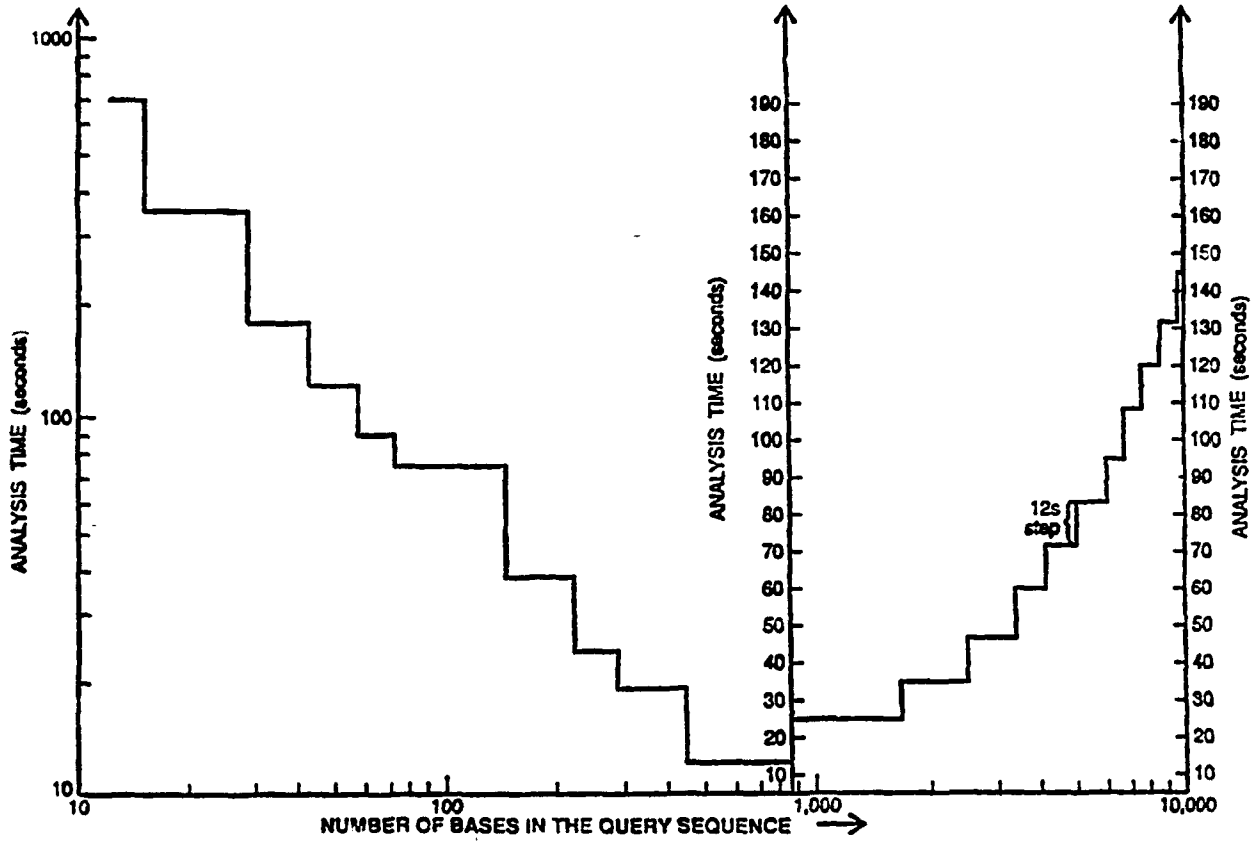

Figure 8: Processing time for the analysis of a 50×10^6 bases database as a function of the number of bases in the query sequence. The left of the figure uses log-log axis and covers query sequences of length 12 to 857. Semi-log axis are more convenient for the right of the figure because the analysis time varies linearly with the length of the query sequence. The abscissa and ordinate are respectively drawn on a logarithmic scale and a linear scale.

3) Query sequences containing between 15 and 28 bases can be analyzed with a repetition of one and a bit rate of 1 MHz.

4) When the query sequences contain more than 429 bases it is possible to use the maximum bit rate of 30 MHz. From there, the analysis time grows linearly with the length of the query sequence and the number of time-shifts that have to be used. Longer query sequences produce many consecutive correlation peaks. For example, a 500-bases query sequence, a 3000-bases query sequence and a 10^5-bases query sequence have respective durations of 116.7 μs, 700 μs and 23333 μs when processed at a bit rate of 30 MHz. One, four and 117 time-shifts have to be tried thus producing total analysis times of 12 seconds, 48 seconds and 22 minutes for a database containing 50×10^6 bases. Performing the same analysis on all 3×10^9 bases of the human genome would take respectively 12 minutes, 48 minutes and 23 hours.

It is possible at this point to compare, for one particular case, the analysis times of the TIC described here with current digital technologies. The analysis of a 3000-bases query sequence in a 50×10^6 database is performed in 48 seconds with the TIC, in 8 minutes with a 80 MIPS mainframe computer and in over 5 hours with a personal computer operating at 25 MHz (see section II). The TIC is then 10 times faster than the 80 MIPS computer and over 375 times faster than the personal computer.

We assume that only one TIC is available to perform the analysis so the various time-shifts have to be tried consecutively. If more than one TIC is available, parallel processing of many 200 μs windows is possible and the analysis time is divided by the number of TICs operating in parallel. If four TICs are available, the analysis times for the examples of the preceding paragraph are 3 minutes, 12 minutes and 5.8 hours. Compact architectures[17-19] have been developed for the construction of TICs and such an approach applied to the reduction of processing time is very feasible.

The next step in the coarse analysis is to transform the time of occurrence of the peaks into location in the database. A rough estimation of the time of occurrence is provided by the frame number where the peaks are found. It is possible to make a more precise determination of the time of occurrence by finding the location of the peak within the frame. At this point, the coarse analysis is complete and a fine analysis of the segments previously identified as interesting has to be performed.

7.0 STRATEGY FOR FINE ANALYSIS

The purpose of fine analysis is to produce a base-by-base comparison between the database and the query sequence. The presence of any discrepancies will be revealed with all the details of these features. The key to fine analysis is to use lower data rates, representations of the bases that are much longer and to perform the analysis only on the segments of interest identified by the coarse analysis. Maximum length pseudorandom sequences containing 127 bits (see Table 3) and an integration time of 127 μs could be used with a data rate of 1 MHz.

Table 3: Long representations of the DNA bases with 127-bits maximum length pseudorandom sequences that are designated [35,p.62] By their octal and their polynomial representations.

	octal representation	polynomial representation
Adenine (A)	203	$x^7 + x + 1$
Cytosine (C)	211	$x^7 + x^3 + 1$
Guanine (G)	217	$x^7 + x^3 + x^2\, x + 1$
Thymine (T)	221	$x^7 + x^4 + 1$
Unknown (N)	235	$x^7 + x^4 + x^3 + x^2 + 1$

When the TIC operates in this mode (see Figure 7), the correlation of the database bases should be synchronized with the bases of the query sequence to optimise the height of the correlation peaks. The controller of the system and the access to the memory containing the query sequence and the database should be designed with enough flexibility to provide the capability to move back and forth in the memory in order to analyse in detail the gaps and discrepancies between the query sequences and the database. The time required to do this analysis is a linear function of the number of bases in the query sequence. As it takes 127 μs to confirm the presence of a particular base at a particular location in the database, a detailed analysis of a 3000-bases and a 10^5-bases query sequence takes less than 2 seconds and 16 seconds respectively. A 20% time overhead is added for the determination of the parameters of gaps and the exact location of the beginning of the query sequence in the database. If there is more than one occurrence of the query sequence in the database, the fine analysis has to repeated each time. The analysis of the reverse complementary strand of a particular query sequence should be treated as a new experiment with a different query sequence and the coarse and fine analysis have to be repeated.

8.0 HIGH PERFORMANCE OPTICAL PROCESSING FOR DNA ANALYSIS

8.1 Introduction

A system dedicated to the analysis of DNA sequences based on a TIC contains conceptually three parts, 1) the optical correlator 2), the Controller and 3), a fast-access, large capacity storage unit for the data from the query sequence and the database. The feasibility of the construction and operation of the first two parts of the system have already been demonstrated at the Defence Research Establishment Ottawa.

8.2 Selection of the Components of a TIC For DNA Analysis

Considering the large amount of data to be processed for DNA analysis, the components of the TIC and the operating procedures should be selected to provide maximum speed of operation. The Bragg cells should have the largest possible transit time τ to maximize the time-delay window 2τ in which to observe a correlation peak between two signals. The bandwidth of the Bragg cells should be maximized also for maximum bit rate operation. In order to avoid producing a distorted correlation peak [34, p.24] the bit rate should not exceed the Bragg cell bandwidth divided by 1.5. The integration time T of the detector array should be minimized to perform the reading operation at a reasonable rate but have enough elements to provide sufficient resolution of the time-delay window and produce accurate determination of the peak position.

A design based on commercially available equipment could include TeO_2 Bragg cells with a 100 μs time aperture and a 50 MHz bandwidth. A maximum data rate of 30 MHz could then be used. Detector arrays with 1000 elements, a read-out rate of 20 MHz and a minimum integration time of 50 μs are available thus generating 20000 correlations per second.

8.3 The Controller

Most aspects of the operation of the TIC are controlled by a computer that we call the Controller. The Controller should maintain an interface with the user that allows 1) the selection of the database to the used for analysis, 2) the input of the query sequence to be analyzed and 3) the display of the results in a format familiar to the user. The Controller is also responsible for the input of the data to the TIC and for the collection of data from the detector array, including pedestal removal and peak detection. It should contain algorithms to define the parameters of operation for the coarse and the fine analysis and should decide on the base representation length, number of repetition of the bits and the data rate to be used. It should also determine the number of repetition required of the analysis for various time-shifted versions of the query sequence.

Work already performed at DREO on a similar system for the processing of communication signals has demonstrated the feasibility of such a system for real-time operation at a data rate of 30 MHz.

8.4 Fast Access Storage for Data Input to the TIC

The storage unit should store billions of DNA bases and prepare the signal representations required by the Bragg cells. A capability to send the data to the TIC at bit rates between 1 and 30 MHz should be available. Flexible, fast access to any part of the data is particularly important for rapid fine analysis. Preliminary consultations led to the conclusion that this task is not trivial but is feasible with existing technology.

9.0 CONCLUSION

Elements of optical data processing and spread-spectrum communication theory have been integrated to present a proposal for the analysis of DNA sequences with an optical TIC. An analysis strategy including a coarse and a fine analysis was developed and the resulting processing times were calculated. It was concluded that TICs could produce a substantial improvement in DNA analysis processing times. Comparison of the expected processing times of a TIC, for a particular case, lead to the conclusion that the TIC could be 10 times faster than a 80 MIPS computer and over 375 times faster than a personal computer. The requirements of an operational system were outlined.

10.0 REFERENCES

[1] W.T. Rhodes, "Optical Information Processing in the 1990's", SPIE vol.1151, Optical Information Processing Systems and Architectures, 1989, p.387-388.

[2] L.Smith and L. Hood, "MAPPING AND SEQUENCING THE HUMAN GENOME: HOW TO PROCEED", BIO/TECHNOLOGY vol.5, Sept. 1987, p.933-939.

[3] R. Lewis,"How Lasers Can Speed Up The Human Genome Project", Photonics Spectra, May 1991, p.72-75.

[4] S.L. Williams, "Imaging the Human Genome", Advanced Imaging, July 1990, p.16-19.

[5] J.F. Hawk, J.C. Martin, D.A Gregory and W.A. Christens-Barry, "Optimum Character Encryption and Extraction for Optical Correlation Techniques",SPIE vol.1151, Optical Information Processing Systems and Architectures, 1989, p.299-306.

[6] W.A. Christens-Barry, J.F. Hawk and J.C. Martin, "Vander Lugt Correlation of DNA Sequence Data", SPIE vol. 1347, Optical Information Processing Systems and Architectures II, 1990, p.221-230.

[7] J.A. Neff, "Some Thoughts on Optical Processing/Computing in the 1990's", SPIE vol.1151, Optical Information Processing Systems and Architectures, 1989, p. 385-386.

[8] M.W. Casseday, N.J. Berg, I.J. Abramovitz and J.N. Lee, "Wide-Band Signal Processing Using the Two-Beam Surface Acoustic Wave Acoustooptic Time Integrating Correlator", IEEE Transactions on Sonic and Ultrasonics, vol.SU-28, no.3, May 1981, p.205-212.

[9] N.J. Berg, I.J. Abramovitz, J.N. Lee and M.W. Casseday," A New Surface-Wave Acousto-Optic Time Integrating Correlator", Appl. Phys. Lett. 36 (4), 15 Feb. 1980, p.256-257.

[10] N.J. Berg, M.W. Casseday, A.N. Filipov and J.M. Pellegrino, " A NEW MULTIFUNCTION ACOUSTO-OPTIC SIGNAL PROCESSOR", 1983 Ultrasonics Symposium, p.454-458.

[11] I.J. Abramovitz, N.J. Berg and M.W. Casseday, "INTERFEROMETRIC SURFACE-WAVE ACOUSTO-OPTIC TIME-INTEGRATING CORRELATORS", 1980 Ultrasonics Symposium, p.483-487.

[12] N.J. Berg, M.W. Casseday, I.J. Abramovitz and J.N. Lee, "Radar and Communication Band Signal Processing Using Time-Integrating Processors", SPIE vol.232, 1980 International Optical Computing Conference, p.101-107.

[13] C.S. Tsai, J.K. Wang and K.Y. Liao, "Acousto-optic Time-Integrating Correlators using Integrated Optic Technology", SPIE vol.180, Real-Time Signal Processing II, 1979, p. 160-163.

[14] M. Varasi, A. Vannucci and S. Reid, "Integrated Acousto-Optic Correlator using the Proton Exchange Technique", SPIE vol.1151, Optical Information Processing Systems and Architectures, 1989, p.457-466.

[15] W.T. Rhodes, "Acousto-Optic Signal Processing: Convolution and Correlation", Proc. IEEE, vol.69, no.1, Jan. 1981, p.65-79.

[16] N. Laouar, J.P. Goedgebuer et R. Ferriere,"CORRELATEUR OPTO-ELECTRONIQUE ANALOGIQUE POUR LE TRAITEMENT EN PARALLELE DE SIGNAUX DE TYPE RADAR", Onzième colloque GRETSI- Nice 1-5 juin 1987, p.693-696.

[17] I.G. Fuss, "Acousto-optic Signal Processor Based on a Mach-Zehnder Interferometer", Appl.Opt., vol.24, no.22, 15 Nov. 1985, p.3866-3871.

[18] D.A.B. Fogg, "A Compact Bulk Acousto-Optic Time Integrating Correlator", Department of Defence of Australia, Technical Report ERL-0323-TR, Nov. 1984.

[19] M.S. Brown,"A Kosters Prism Time-Integrating Acousto-Optic Correlator", J. Phys, E:Sci. Instrum. 21 (1988) 192-194.

[20] N. Brousseau and J.W.A. Salt, "Design and Implementation of a Time-Integrating Correlator Using Bulk Acousto-Optics Interaction", Defence Research Establishment Ottawa Technical Note 86-25, Sept. 1986.

[21] N. J. Berg and J.N. Lee, "Acousto-Optic Signal Processing: Theory and Implementation", Marcel Dekker Inc. New York and Basel, 1983.

[22] R.A. Sprague and C.L. Koliopoulos, "Time Integrating Acoustooptic Correlator", Appl.Opt., vol.15, no.1, Jan 1976, p.89-92.

[23] C.C. Lee, K.Y. Liao and C.S. Tsai, "Acousto-Optic Time-Integrating Correlator Using Hybrid Integrated Optics", 1982 IEEE Ultrasonics Symposium p. 405-407.

[24] G. Silbershatz and D. Casasent, "Hybrid Time and Space Integrating Processors for Spread Spectrum Applications", Appl. Opt., vol.22, no.14, 15 July 1983, p.2095-2103.

[25] D. Casasent, "General Time-, Space-, and Frequency -Multiplexed Acoustooptic Correlator", Appl. Opt., vol.24, no.17, 1 Sept. 1985., p. 2884-2888.

[26] F.B. Rotz, "Time-Integrating Optical Correlator", SPIE vol.202, Active Optical Devices, 1979, p.163-169.

[27] P. Kellman, "Time-Integrating Optical Signal Processing", Stanford University, Dept. of Electrical Engineering, Ph.D. dissertation, June 1979.

[28] A.P. Goutzoulis and B.V.K. Vijaya Kumar, "Optimum Time-Integrating Acousto-Optic Correlator for Binary Codes", Optics Communications, vol.48, no.6, 15 Jan. 1984, p.393-397.

[29] I.D. Bondarenko, A.A. Vetrov and Y.V. Popov, "Analysis of the Errors of the Signal Processing Channel of an Acousto-Optic Correlator", Sov. J. Opt. Technol.56 (6), June 1989, p.346-349.

[30] D. Casasent, A. Goutzoulis and V.K. Vijaya Kumar, "Time-Integrating Acoustooptic Correlator: Error Source Modelling", Appl. Opt., vol.23, no.18, 15 Sept. 1984, p.3130-3137.

[31] B.V.K. Vijaya Kumar and J.M. Connelly, "Binarization Effects in Acousto-Optic Correlators", SPIE vol.1347 Optical Information Processing Systems and Architectures II, 1990, p.112-122.

[32] J.B. Goodell, "Optical Design Considerations for Acousto-Optic Systems", SPIE vol.936, Advances in Optical Information Processing III, 1988, p.22-28.

[33] A. Goutzoulis, D. Casasent and B.V.K. Vijaya Kumar, "Detector Effects on Time-Integrating Correlators Performance", Appl. Opt., vol.24, no.8, 15 April 1985, p.1224-1233.

[34] R.C. Dixon, "Spread Spectrum Systems", John Wiley & Sons, 1984.

[35] S.W. Golomb, "Shift Register Sequences", Aegean Park Press, Revised Edition 1982.

SECURITY CLASSIFICATION OF FORM
(highest classification of Title, Abstract, Keywords)

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Establishment sponsoring a contractor's report, or tasking agency, are entered in section 8.)
NATIONAL DEFENCE
DEFENCE RESEARCH ESTABLISHMENT OTTAWA
SHIRLEY BAY, OTTAWA, ONTARIO K1A 0K2 CANADA

2. SECURITY CLASSIFICATION (overall security classification of the document including special warning terms if applicable)
UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C or U) in parentheses after the title.)
ANALYSIS OF DNA SEQUENCES BY AN OPTICAL TIME-INTEGRATING CORRELATOR: PROPOSAL (U)

4. AUTHORS (Last name, first name, middle initial)
BROUSSEAU, N. AND BROUSSEAU, R.

5. DATE OF PUBLICATION (month and year of publication of document)
NOVEMBER 1991

6a. NO. OF PAGES (total containing information. Include Annexes, Appendices, etc.)
31

6b. NO. OF REFS (total cited in document)
35

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)
DREO TECHNICAL NOTE

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include the address.)
NATIONAL DEFENCE
DEFENCE RESEARCH ESTABLISHMENT OTTAWA
SHIRLEY BAY, OTTAWA, ONTARIO K1A 0K2 CANADA

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant)
041LQ11

9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written)

10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique to this document.)
DREO TECHNICAL NOTE 91-33

10b. OTHER DOCUMENT NOS. (Any other numbers which may be assigned this document either by the originator or by the sponsor)

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Distribution limited to defence departments and defence contractors; further distribution only as approved
( ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved
( ) Distribution limited to government departments and agencies; further distribution only as approved
( ) Distribution limited to defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availabilty (11). However, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)

UNCLASSIFIED

SECURITY CLASSIFICATION OF FORM

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both offical languages unless the text is bilingual).

(U) This technical note presents a proposal to perform the analysis of DND sequences with an analog optical computer. The DNA analysis involves the computation of massive amount of correlations. A time-integrating correlator is an ideal tool to perform that processing at very fast speed. A design based on commercially available equipment is presented together with a comparison of the processing time of the system with conventional computer technology. The improvement in speed is of orders of magnitude and sufficient to make the analysis of the whole human genome a tractable enterprise. An overview of the technology already available for such a project is presented together with an outline of the areas that need more development.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

DNA
TIME-INTEGRATING CORRELATOR
OPTICAL DATA PROCESSING

UNCLASSIFIED

SECURITY CLASSIFICATION OF FORM